ECONOMIC PREDICTIONS WITH BIG DATA: THE ILLUSION OF SPARSITY

DOMENICO GIANNONE, MICHELE LENZA, AND GIORGIO E. PRIMICERI

Abstract. We compare sparse and dense representations of predictive models in macroeconomics, microeconomics and finance. To deal with a large number of possible predictors, we specify a "spike-and-slab" prior that allows for both variable selection and shrinkage. The posterior distribution does not typically concentrate on a single sparse or dense model, but on a wide set of models. A clearer pattern of sparsity can only emerge when models of very low dimension are strongly favored a priori.

1. Introduction

The recent availability of large datasets, combined with advances in the fields of statistics, machine learning and econometrics, have generated interest in predictive models with many possible predictors. In these cases, standard techniques such as ordinary least squares, maximum likelihood, or Bayesian inference with uninformative priors perform poorly, since the proliferation of regressors magnifies estimation uncertainty and produces inaccurate out-of-sample predictions. As a consequence, inference methods aimed at dealing with the curse of dimensionality have become increasingly popular.

In very general terms, these methodologies can be divided in two broad classes. *Dense*-modeling techniques recognize that all possible explanatory variables might be important for prediction, although their individual impact might be small. Factor analysis or ridge regressions are standard examples of dense statistical modeling (Pearson, 1901, Spearman, 1904, Lawley and Maxwell, 1963, Tikhonov, 1963, Hoerl and Kennard, 1970, Leamer, 1973; see also Stock and Watson, 2002a,b and De Mol et al., 2008 for big data applications of these techniques in economics). At the opposite side of the spectrum, *sparse*-modeling techniques focus on selecting a small set of explanatory variables with the highest predictive power,

Date: First version: March 2017. This version: July 2017.
We thank Mike West, as well as seminar and conference participants for comments and suggestions. The views expressed in this paper are those of the authors and are not necessarily reflective of views at the European Central Bank, the Eurosystem, the Federal Reserve Bank of New York, or the Federal Reserve System.

out of a much larger pool of regressors. For instance, the popular lasso belongs to this class of estimators that produce sparse representations of predictive models (Tibshirani, 1996, Hastie et al., 2015; see also Belloni et al., 2011 for a recent survey and examples of big data applications of these methodologies in economics).

In this paper, we ask whether sparse modeling is a good approach to predictive problems in economics, compared to dense modeling. Observe that standard variable-selection techniques are guaranteed to consistently recover the pattern of sparsity only if the true model is actually sparse, or approximately so. Therefore, lasso-based methods cannot be used to answer the question whether sparsity leads to good predictions because sparsity is essentially assumed. Moreover, if the true data-generating process is dense, it is unclear what sparse estimators deliver. They might select a small set of explanatory variables simply because sparsity provides a way of reducing estimation uncertainty, overcoming the curse of dimensionality and improving prediction accuracy, and not because the true model is actually sparse.

Instead, we propose to study the suitability of sparse predictive models in a framework that allows for sparsity, but does not assume it. We specify a so-called "spike-and-slab" prior for the coefficients of a linear predictive model, in the spirit of Mitchell and Beauchamp (1988). This prior states that regression coefficients can be non-zero with a certain probability q. We refer to this hyperparameter as the probability of inclusion. When a coefficient is not zero, it is modeled as a draw from a Gaussian distribution. The variance of this density is scaled by the hyperparameter γ^2, which thus controls the degree of shrinkage when a predictor is included. The higher γ^2, the higher the prior variance, the less shrinkage is performed. In sum, our model has two key ingredients. First, it shrinks the non-zero coefficients towards zero, to reduce estimation uncertainty and give a chance to large-dimensional models. Second, it treats shrinkage and variable selection separately, as they are controlled by different hyperparameters, γ^2 and q. These hyperparameters are treated as unknown and we conduct inference on them.

We estimate our model on six datasets that have been used in the literature for macroeconomic, finance and microeconomic predictions with large information. In our macroeconomic applications, we investigate the predictability of economic activity in the US, and the determinants of economic growth in a cross-section of countries. In finance, we study the predictability of the US equity premium, and the factors that explain the cross-sectional

variation of US stock returns. Finally, in our microeconomic analyses, we investigate the factors behind the decline in the crime rate in a cross-section of US states, and the determinants of rulings in the matter of government takings of private property in US judicial circuits.

Our Bayesian inferential method delivers three main results. First, we characterize the marginal posterior distribution of the probability q of inclusion for each predictor. Only in one case, the first microeconomic application, this posterior is concentrated around very low values of q. In all other applications, larger values of q are more likely, suggesting that including more than a handful of predictors is preferable in order to achieve a good forecasting accuracy. Second, the joint posterior distribution of q and γ^2 typically exhibits a clear negative correlation: the higher the probability of including each predictor, the higher the likelihood of needing more shrinkage. This intuitive finding explains why larger-scale models forecast well in our framework.

Third, while the appropriate degree of shrinkage and model size are quite well identified, the data are much less informative about the specific set of predictors that should be included in the model. Put differently, model uncertainty is pervasive, and the best prediction accuracy is not achieved by a single model, but by averaging many models with rather different sets of predictors. As a consequence, it is difficult to characterize the resulting representation as sparse, which explains the reference to the "illusion of sparsity" in our title. According to our results, a clearer pattern of sparsity can only emerge when the researcher has a strong a-priori bias in favor of predictive models with a small number of regressors.

On a separate note, an important last point to emphasize is that the definition of sparsity is not invariant to transformations of the regressors. For example, consider a model in which only the first principal component of the explanatory variables matters for prediction. Such a model is sparse in the rotated space of the predictors corresponding to the principal components. It is instead dense in the untransformed, or "natural" space of the original regressors, since the first principal component combines all of them. This paper studies the issue of sparsity versus density in this natural space of the untransformed regressors. There are a number of reasons that motivate this focus. The first one is comparability with the literature on lasso and variable selection, which typically assumes the existence of a sparse representation in terms of the original predictors. Second, analyzing sparsity patterns in this

natural space is usually considered more interesting from an economic perspective because it is easier, and thus more tempting, to attach economic interpretations to models with few untransformed predictors. Third, for any model, it is always possible to construct a rotated space of the predictors a posteriori, with respect to which the representation is sparse. Therefore, the question of sparsity versus density is meaningful only with respect to spaces that are chosen a priori—such as that of the original regressors or of a priori transformations of them—and do not depend on the response variable and the design matrix.

The rest of the paper is organized as follows. Section 2 describes the details of our prediction model. Section 3 and 4 illustrate the six economic applications and the main estimation results.

2. Model

We consider the following linear model to predict a response variable y_t,

$$y_t = u_t'\phi + x_t'\beta + \varepsilon_t, \tag{2.1}$$

where ε_t is an i.i.d. Gaussian error term with zero mean and variance equal to σ^2, and u_t and x_t are two vectors of regressors of dimensions l and k respectively, with typically $k \gg l$, and whose variance has been normalized to one. Without loss of generality, the vector u_t represents the set of explanatory variables that a researcher always wants to include in the model, for instance a constant term. Therefore, the corresponding regression coefficients ϕ are never identically zero. Instead, the variables in x_t represent possibly, but not necessarily useful predictors of y_t, since some elements of β might be zero. When this is the case, we say that (2.1) admits a sparse representation.

To capture these ideas, and address the question of whether sparse representations of economic predictive models fit the data well, we specify the following prior distribution for the unknown coefficients $\left(\sigma^2, \phi, \beta\right)$,

$$p\left(\sigma^2\right) \propto \frac{1}{\sigma^2}$$

$$\phi \sim \text{flat}$$

$$\beta_i|\sigma^2, \gamma, q \sim \begin{cases} N\left(0, \sigma^2\gamma^2\right) & \text{with pr. } q \\ 0 & \text{with pr. } 1-q \end{cases} \qquad i = 1, ..., k$$

The priors for the low dimensional parameters ϕ and σ^2 are rather standard, and designed to be uninformative. Instead, the elements of the vector β are either zero, with probability $1-q$, or normally distributed with the same variance, given the standardization of the regressors. The hyperparameter γ^2 plays a crucial role since it controls the variance of this Gaussian density, and thus the degree of shrinkage when a regressor is included in the model. Without the possibility of shrinkage, the only way to improve prediction accuracy and avoid overfitting in large-dimensional models would be through variable selection. As a consequence, sparsity would emerge almost by construction.

A similar way to describe our prior for β would be to say that $\beta_i|\sigma^2,\gamma,q \sim N\left(0,\sigma^2\gamma^2\nu_i\right)$ for $i=1,...,k$, with $\nu_i \sim Bernoulli\left(q\right)$. This formulation is useful because it highlights the relation between our model and some alternative specifications adopted in the literature on dimension reduction and sparse-signal detection. For example, the Bayesian ridge regression corresponds to simply setting $q=1$. The Bayesian lasso and horseshoe methods can instead be obtained by replacing the Bernoulli distribution for ν_i with an exponential or a half-Cauchy density respectively (Park and Casella, 2008 and Carvalho et al., 2010). None of these alternative priors, however, admit a truly sparse representation of (2.1) with positive probability.

Our prior on β belongs to the so-called "spike-and-slab" class, initially proposed by Mitchell and Beauchamp (1988) to perform variable selection and find sparse representations of linear regression models. Differently from them, however, the "slab" part of our prior is not a uniform density but a Gaussian, as in George and McCulloch (1993, 1997), and Ishwaran and Rao (2005). In addition, relative to most variants of the spike-and-slab prior adopted in the literature on variable selection, we treat the hyperparameters q and γ^2 as unknown and evaluate their posterior distribution, along the lines of George and Foster (2000) and Liang et al. (2008). They are crucial objects of interest for our analysis of sparsity patterns.

To specify a hyperprior on q and γ^2, we define the mapping $R^2\left(\gamma^2,q\right) \equiv \frac{qk\gamma^2 var(x)}{qk\gamma^2 var(x)+1}$ and place the following independent priors on q and R^2:

$$q \sim Beta\left(a,b\right)$$

$$R^2 \sim Beta\left(A,B\right).$$

The marginal prior for q is a Beta distribution, with support $[0,1]$, and shape coefficients a and b. In most of our empirical applications, we will work with $a = b = 1$, which corresponds to a uniform prior. We will also experiment with prior densities skewed to the right, which assign a much higher probability to models with low values of q and a limited number of regressors. Turning to γ^2, it is difficult to elicit a prior directly on this hyperparameter. The function $R^2\left(\gamma^2, q\right)$, instead, has an intuitive interpretation as $var\left(x_t'\beta|\gamma^2, q, \sigma^2\right)/var\left(x_t'\beta + \varepsilon_t|\gamma^2, q, \sigma^2\right)$, i.e. as the share of the variance of y_t due to the $x_t'\beta$ term relative to the error. We model this ratio as a Beta distribution with shape coefficients A and B, and base our inference on the uninformative case with $A = B = 1$. The appeal of this hyperprior is that it can be used for models of possibly very different size, because it has the interpretation of a prior on the R^2 of the regression. Another attractive feature is that it implies a negative prior correlation between q and γ^2, and is thus agnostic about whether the curse of dimensionality should be dealt with variable selection or shrinkage. We will return to this point in section 4, when discussing our posterior results.

3. Economic Applications

We estimate the previous model on six popular "big datasets" that have been used for predictive analyses in macroeconomics, finance and microeconomics. In our macroeconomic applications, we investigate the predictability of economic activity in the US (macro 1) and the determinants of economic growth in a cross-section of countries (macro 2). In finance, we study the predictability of the US equity premium (finance 1) and the factors that explain the cross-sectional variation in expected US stock returns (finance 2). Finally, in our microeconomic applications, we investigate the effects of legalized abortion on crime in a cross-section of US states (micro 1) and the determinants of rulings in the matter of government takings of private property in US judicial circuits (micro 2). Table 1 summarizes the data used in the analysis. A more detailed description is provided in the text below.

3.1. **Macro 1: Macroeconomic forecasting using many predictors.**

In this application, we study the importance of large information to forecast US economic activity, an issue investigated by a large body of time-series research in the last decade. We use a popular large dataset originally developed for macroeconomic predictions with principal components by Stock and Watson (2002a,b), and extensively used to assess the forecasting performance of alternative big-data methodologies. The variable to predict is the monthly

TABLE 1. Description of the datasets.

	Dependent variable	Possible predictors	Sample
Macro 1	Monthly growth rate of US industrial production	130 lagged macroeconomic indicators	659 monthly time-series observations, from February 1960 to December 2014
Macro 2	Average growth rate of GDP over the sample 1960-1985	60 socio-economic, institutional and geographical characteristics, measured at pre-60s value	90 cross-sectional country observations
Finance 1	US equity premium (S&P 500)	16 lagged financial and macroeconomic indicators	58 annual time-series observations, from 1948 to 2015
Finance 2	Stock returns of US firms	144 dummies classifying stock as very low, low, high or very high in terms of 36 lagged characteristics	1400k panel observations for an average of 2250 stocks over a span of 624 months, from July 1963 to June 2015
Micro 1	Per-capita crime (murder) rates	Effective abortion rate and 284 controls including possible covariate of crime and their transformations	576 panel observations for 48 US states over a span of 144 months, from January 1986 to December 1997
Micro 2	Number of pro-plaintiff eminent domain decisions in a specific circuit and in a specific year	Characteristics of judicial panels capturing aspects related to gender, race, religion, political affiliation, education and professional history of the judges, together with some interactions among the latter, for a total of 138 regressors	312 panel circuit/year observations, from 1975 to 2008

growth rate of US industrial production, and the dataset consists of 130 possible predictors, including a variety of macroeconomic monthly indicators, such as measures of output, income, consumption, orders, surveys, labor market variables, house prices, consumer and producer prices, money, credit and asset prices. The sample ranges from February 1960 to December 2014, and all the data have been transformed to obtain stationarity, as in the work of Stock and Watson. The version of the dataset that we use is available at FRED-MD, and is regularly updated through the Federal Reserve Economic Data (FRED), a database maintained by the Research division of the Federal Reserve Bank of St. Louis (McCracken and Ng, 2016).

3.2. **Macro 2: The determinants of economic growth in a cross-section of countries.** The seminal paper by Barro (1991) initiated a debate on the cross-country determinants of long-term economic growth. Since then, the literature has proposed a wide range of possible predictors of long-term growth, most of which have been collected in the dataset constructed by Barro and Lee (1994). As in Belloni et al. (2011), we use this dataset to explain the average growth rate of GDP between 1960 and 1985 across countries. The database includes data for 90 countries and 60 potential predictors, corresponding to the pre-1960 value of several measures of socio-economic, institutional and geographical characteristics. The logarithm of a country's GDP in 1960 is always included as a regressor in the model.[1]

3.3. **Finance 1: Equity premium prediction.** Following a large body of empirical work, in our first finance application we study the predictability of US aggregate stock returns, using the database described in Welch and Goyal (2008). More specifically, the dependent variable is the US equity premium, defined as the difference between the return on the S&P 500 index and the 1-month Treasury bill rate. As possible predictors, the dataset includes sixteen lagged variables deemed as relevant in previous studies, such as stock characteristics (the dividend-price ratio, the dividend yield, the earning-price ratio, the dividend-payout ratio, the stock variance, the book-to-market ratio for the Dow Jones Industrial Average, the net equity expansion and the percent equity issuing), interest rate related measures (the Treasury bill, the long-term yield, the long-term return, the term spread, the default-yield spread and the defaults-return spread) and some macroeconomic indicators (inflation and the investment to capital ratio). The data are aggregated annually, from 1948 to 2015.[2]

3.4. **Finance 2: Explaining the cross section of expected returns.** Despite the simple characterization of equity returns provided by the workhorse CAPM model, the empirical finance literature has discovered many factors that can explain the cross-section of expected asset returns. The recent survey of Harvey et al. (2016) identifies about 300 of these factors. Following this tradition, in this application we study the predictability of the cross-section of US stock returns, based on the dataset of Freyberger et al. (2017).[3]

[1]We have downloaded the dataset from the replication material of Belloni et al. (2011), who consider exactly the same application.

[2]We use an updated version of the database downloaded from the webpage of Amit Goyal.

[3]We thank Joachim Freyberger, Andreas Neuhierl and Michael Weber for sharing the database used in their paper.

Our dependent variable is the monthly stock return of firms incorporated in the US and trading on NYSE, Amex and Nasdaq, from July 1963 to June 2015, which results in about 1,400k observations. The set of potential regressors are constructed using (the lagged value of) 36 firms and stocks characteristics, such as market capitalization, the return on assets and equity, the book-to-market ratio, the price-dividend ratio, etc... Inspired by Freyberger et al. (2017), for each of these characteristics we create four dummy variables that take the value of one if the firm belongs to the first, second, fourth or fifth quintile of the distribution within each month, respectively.[4] This results into 144 possible regressors.

3.5. **Micro 1: Understanding the decline in crime rates in US states in the 1990s.** Using US state-level data, Donohue and Levitt (2001) find a strong relationship between the legalization of abortion following the Roe vs Wade trial in 1973, and the subsequent decrease in crime rates. Their dependent variable is the change in log per-capita murder rates between 1985 and the 1997 across US states. This variable is regressed on a measure of the effective abortion rate (which is always included as a predictor) and a set of controls. The latter capture other possible factors contributing to the behavior of crime rates, such as the number of police officers per 1000 residents, the number of prisoners per 1000 residents, personal income per capita, the unemployment rate, the level of public assistance payments to families with dependent children, beer consumption per capita, and a variable capturing the shall-issue concealed carry laws. In addition, as in Belloni et al. (2014), we expand the set of original controls of Donohue and Levitt (2001), by including these variables in levels, in differences, in squared-differences, their cross-products, their initial conditions and their interaction with linear and squared time trends. This extended database includes 284 variables, each with 576 observations relating to 48 states for 12 years.[5]

3.6. **Micro 2: The determinants of government takings of private property in US judicial circuits.** Chen and Yeh (2012) investigate the economic impact of eminent domain in the US, i.e. the right of the government to expropriate private property for public use. To address the possible endogeneity problem, they propose to instrument judicial decisions on eminent domain using the characteristics of randomly assigned appellate courts judges.

[4]The dummy variable equal to one if the firm belongs to the third quintile is excluded for collinearity reasons.

[5]We downloaded the data from the replication material of Belloni et al. (2014), who consider exactly the same application.

For example, it is observed that circuit/years with a higher proportion of black judges are associated with more pro-government decisions in eminent domain cases. We follow Belloni et al. (2012) and estimate the first stage of this instrumental-variable model, by regressing the number of pro-plaintiff appellate decisions in takings law rulings from 1975 to 2008 across circuits on a set of characteristics of the judicial panels such as gender, race, religion, political affiliation, education and professional history of the judges. As in Belloni et al. (2012), we augment the original set of instruments with many interaction variables, resulting into 138 regressors. The sample size (circuit/year units) consists of 312 observations.[6]

4. Exploring the Posterior

In this section, we discuss some properties of the posterior distribution of our model, when estimated using the six datasets illustrated in the previous section. The results we report are based on a uniform prior on q and R^2, the probability of inclusion and the share of the variance of y_t explained by the predictors. We will also experiment with an informative prior centered on low values of q.

4.1. Positive correlation between probability of inclusion and degree of shrinkage.

Our inference method allows us to characterize the joint distribution of the hyperparameters q and γ^2, i.e. the probability of inclusion and the prior variance of the coefficients of the included predictors. The left panels of figures 4.1 and 4.2 summarize the shape of the prior of these two hyperparameters in our six empirical applications, with lighter areas corresponding to higher density regions.[7] We present the joint density of q and $log(\gamma)$, instead of q and γ^2, to interpret the horizontal axis more easily in terms of percent deviations. As we noted in section 2, our flat prior on q and R^2 implies a negative correlation between q and $log(\gamma)$, reflecting the sensible prior belief that probability of inclusion and shrinkage are complements.

The right panels of figures 4.1 and 4.2 show the posteriors of q and $log(\gamma)$. These densities are typically much more concentrated than the corresponding prior, exhibiting an even sharper negative correlation: the lower (higher) the probability of including a predictor

[6] We have downloaded the dataset from the replication material of Belloni et al. (2012), who consider exactly the same application.

[7] The darkness of these plots is also adjusted to correctly capture the relative scale of prior and corresponding posterior.

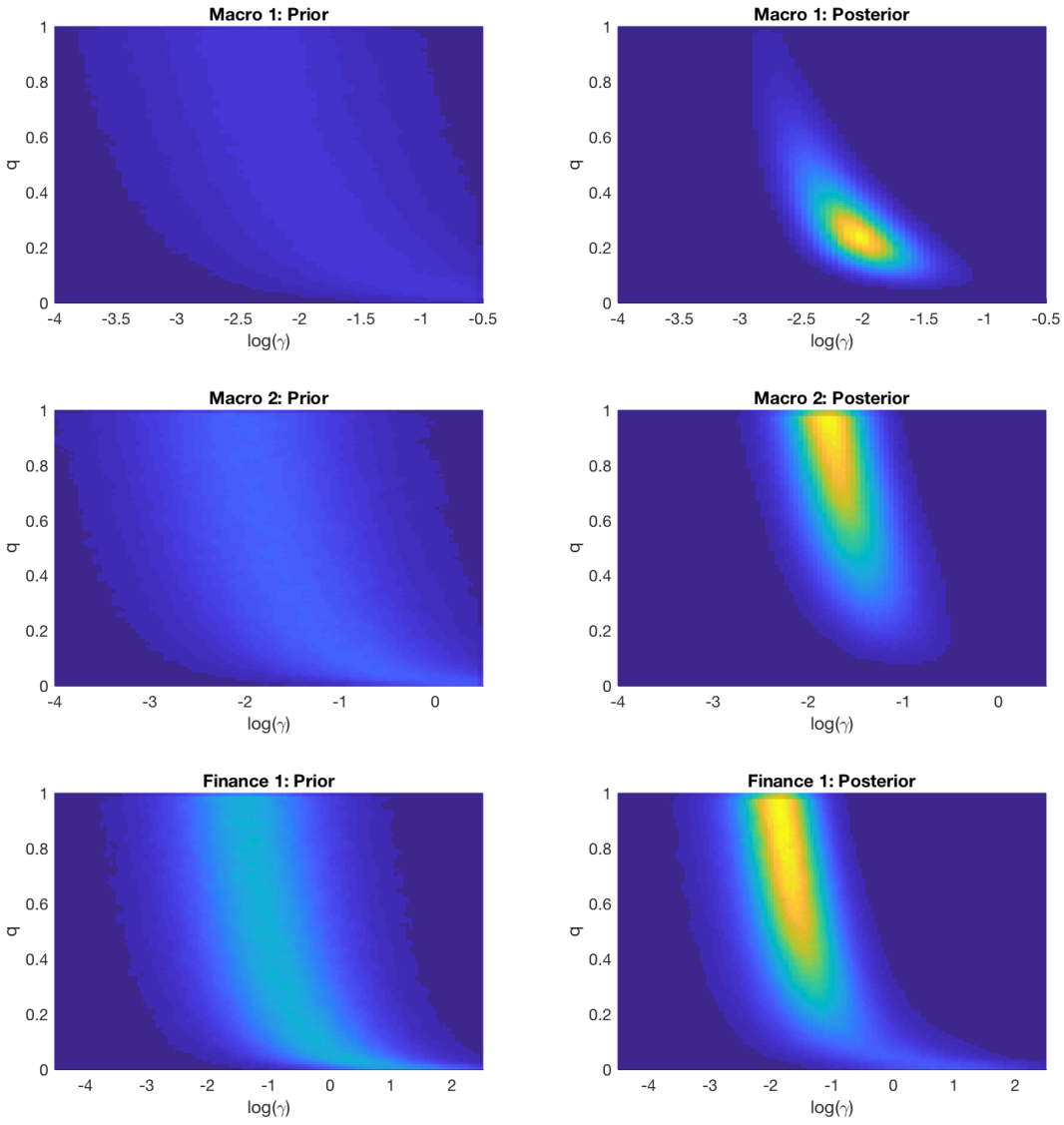

FIGURE 4.1. Joint prior and posterior densities of q and $log\,(\gamma)$ in the macro 1, macro 2 and finance 1 applications (best viewed in color).

and the overall model size, the higher (lower) the prior variance of the coefficients of the predictors included in the model. In other words, larger-scale models need more shrinkage to fit the data well, while models with a low degree of shrinkage require the selection of a smaller number of explanatory variables.

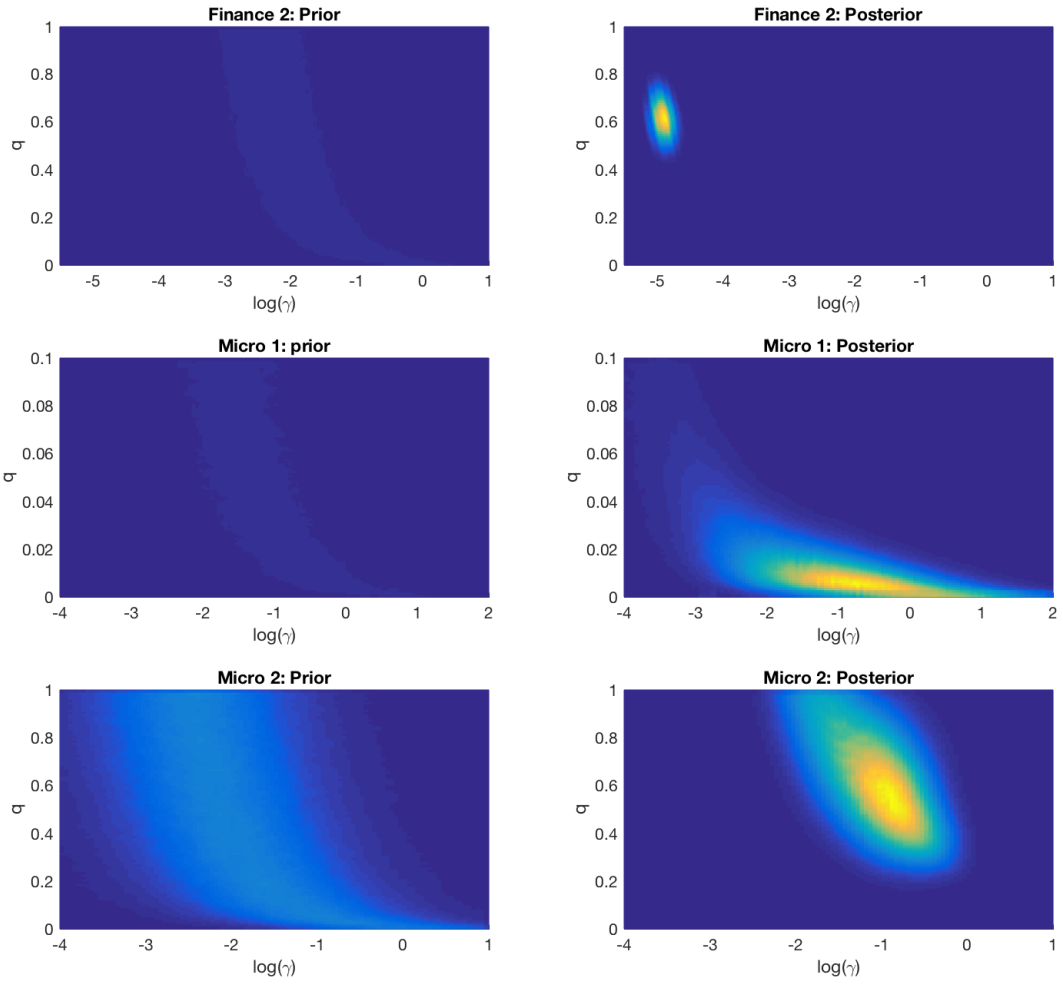

FIGURE 4.2. Joint prior and posterior densities of q and $log\,(\gamma)$ in the finance 2, micro 1 and micro 2 applications (best viewed in color).

While this result should not be particularly surprising, its important implication is that dense modeling approaches might overstate the degree of shrinkage needed to achieve good fit. Similarly, variable selection techniques that do not explicitly allow for shrinkage might artificially recover sparse representations of a model, simply as a device to reduce estimation error. Our findings indicate that these extreme strategies might perhaps be appropriate only for our micro-1 application, given that its posterior in figure 4.2 is tightly concentrated around extremely low values of q. More generally, however, our results suggest that the best prediction models are those that optimally combine probability of inclusion and shrinkage.

4.2. **Probability of inclusion and out-of-sample predictive accuracy.** What is then the appropriate probability of inclusion, considering that models with different sizes require differential shrinkage? To answer this question, figure 4.3 plots the marginal posterior of q, obtained by integrating out γ^2 from the joint posterior distribution of figures 4.1 and 4.2. Notice that the densities in figure 4.3 behave quite differently across applications. For example, the finance-1 data seem to contain little information about model size, since the posterior of q does not deviate much from its uniform prior. The macro-2 application favors more strongly models including the full set of predictors. At the opposite extreme, micro 1 is the only application in which the highest posterior density is concentrated on low values of q, suggesting that the model might be sparse. Macro 1, finance 2 and micro 2, instead, represent intermediate cases, in which the posterior of q is nicely shaped and peaks at an interior point of the $[0, 1]$ interval.

What are the implications of the results of figure 4.3 in terms of goodness of fit? For example, in our macro-1 application, the posterior mode of q is around 0.25. Would working with a dense model ($q = 1$) lead to a substantial deterioration of fit, compared to a model with $q \approx 0.25$? The easiest way to address these questions is to interpret the marginal posterior of q as a measure of out-of-sample predictive accuracy. Observe that, under a flat prior on q, its posterior is proportional to the likelihood,

$$p\left(q|y\right) \propto p\left(y|q\right) = \prod_{t=1}^{T} p\left(y_t|y^{t-1}, q\right), \tag{4.1}$$

where the equality follows from the usual decomposition of a joint density into the product of marginal and conditional densities, and we are omitting the regressors u and x from the conditioning sets to streamline the notation. Expression (4.1) makes clear that the posterior of q is proportional to a product of predictive scores: it corresponds to the probability density that models with different values of q generate zero *out-of-sample* prediction errors. As a consequence, the choice of models with high $p\left(q|y\right)$ can be interpreted as a model selection strategy based on cross validation.

To quantify the variation in predictive accuracy across models with different q's, figure 4.4 plots the function $\frac{1}{T}\left[\log p\left(y|q\right) - \log p\left(y|q^*\right)\right]$ in our six economic applications. This expression corresponds to the average log-predictive score relative to the model with the best fitting q. For instance, in our macro-2 application, the best predictive model is the dense

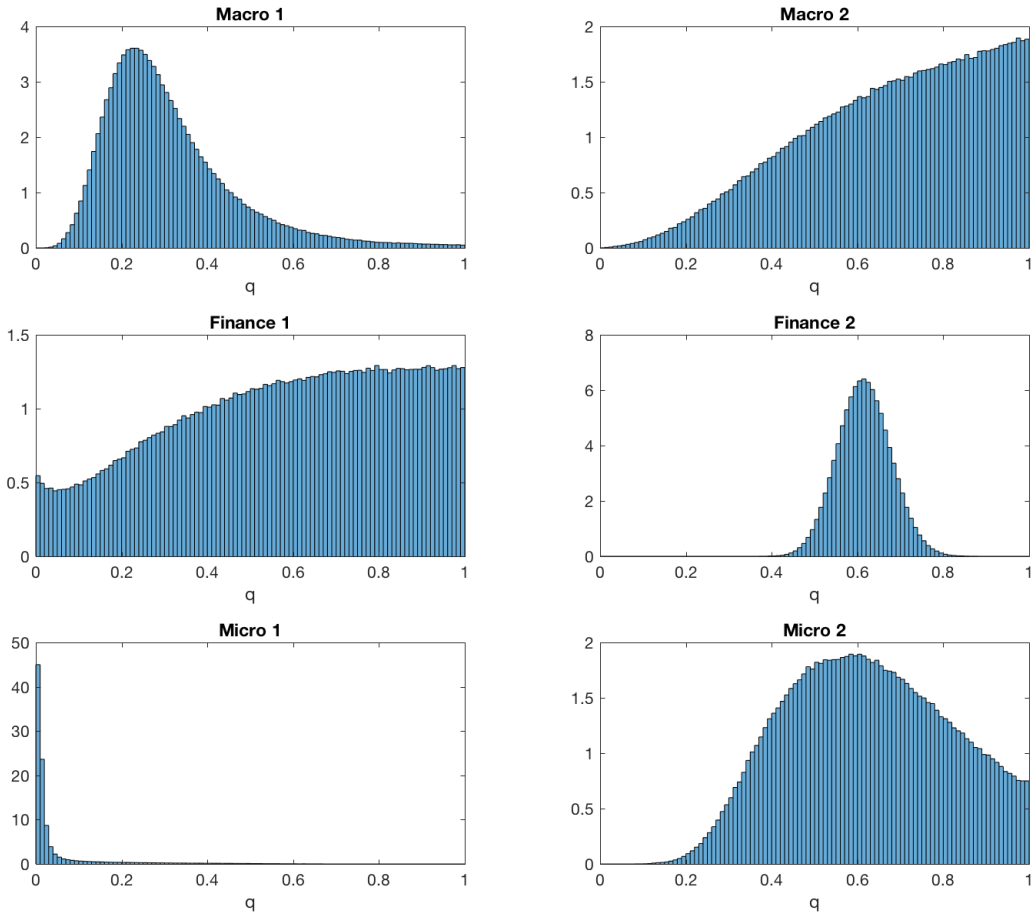

FIGURE 4.3. Posterior density of q.

one, and the top-right panel of figure 4.4 summarizes the average percentage deterioration in the log-predictive score when q declines. As for macro 1, values close to the actual realizations of y are on average $2/3$ of a percent more likely according to a model with $q \approx 0.25$ relative to one with $q = 1$. Low values of q lead to a similar deterioration in forecasting accuracy. For example, a model with $q \approx 0.05$ has approximately the same average log-predictive score of the fully dense model with $q = 1$. When cumulated across the 659 observations, these differences in predictive accuracy can become substantial. The remaining panels in figure 4.4 can be interpreted in a similar way.

4.3. **Patterns of sparsity.** Given these results, we now ask more explicitly to what extent sparse predictive models are appropriate in each of our six applications. The answer to this

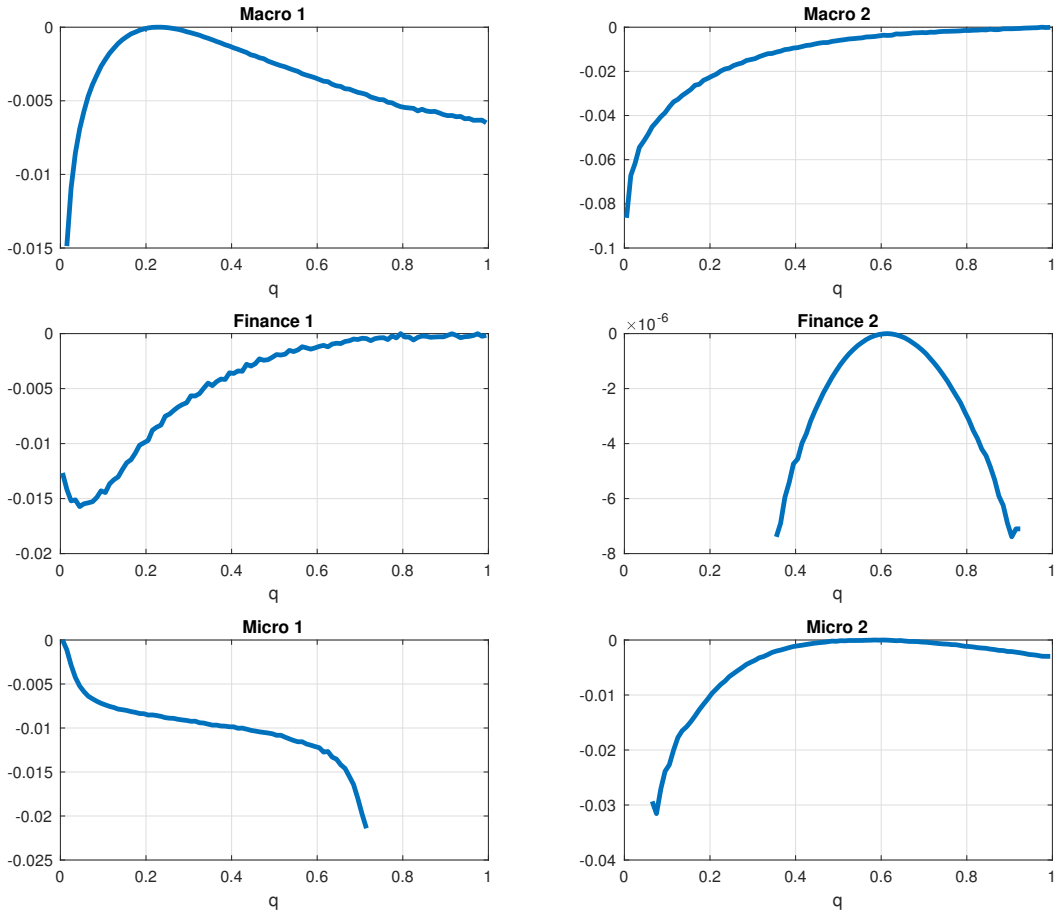

FIGURE 4.4. Average log-predictive score relative to best fitting model.

question turns out to be a bit subtler than one might think based on the information contained in figures 4.3 and 4.4. We will see that a clear pattern of sparsity emerges only in our micro-1 application. Instead, perhaps surprisingly, our posterior results do not support a characterization of the best-fitting predictive models as sparse, not even when the posterior density of q is concentrated around values smaller than 1, as in the macro-1, finance-2 and micro-2 case.

To illustrate this point, figure 4.5 plots the posterior probabilities of inclusion of each predictor. In the "heat maps" of this figure, each vertical stripe corresponds to a possible predictor, and darker shades denote higher probabilities of inclusion. Notice that the probability of inclusion of a single regressor might deviate considerably from q, although the

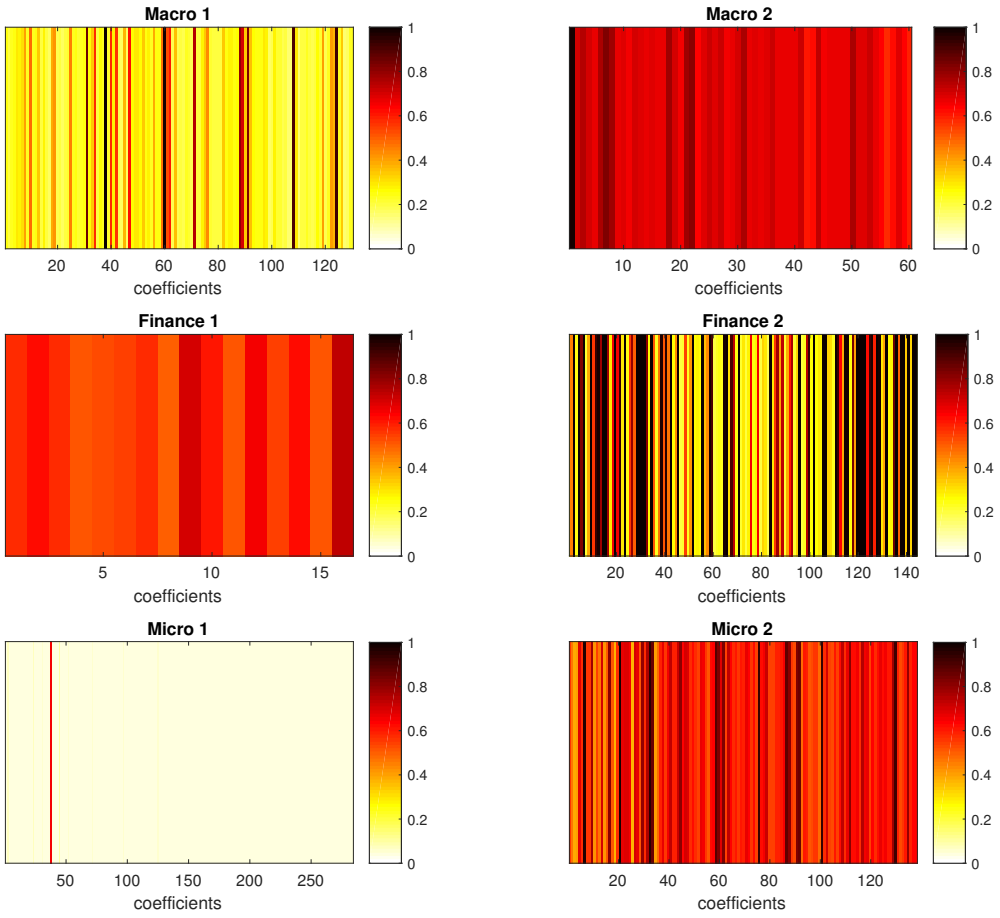

FIGURE 4.5. Heat map of the probabilities of inclusion of each predictor (best viewed in color).

average probability of inclusion across regressors should be consistent with the posterior of q.

The most straightforward subplot to interpret is the one corresponding to the micro-1 application. This is a truly sparse model, in which the 39th regressor is selected 65 percent of the times. The 46th regressor is also sometimes selected, about 10 percent of the times, although this is more difficult to see from the plot. All other predictors are included in the model much more rarely.

The most important message of figure 4.5, however, is that the remaining five applications do not exhibit a distinct pattern of sparsity, in the sense that none of the predictors appear to

be systematically excluded. This finding was probably expected for macro 2 and finance 1, since the posterior of q peaks around very high values in these two applications. The absence of clear sparsity patterns, however, should be more surprising when the posterior of q has most of its mass on lower values. For example, let us consider the case of macro 1, in which the best fitting models are those with q around 0.25, according to figure 4.3. This value of q, however, does not necessarily imply that the most accurate model includes 32 specific predictors (25 percent of the 130 possible regressors) and excludes all the others. If this was the case, the first panel of figure 4.5 would show many near-white stripes corresponding to the predictors that are systematically excluded. Instead, there seems to be a lot of uncertainty about whether certain predictors should be included in the model, which results into their selection only in a subset of the posterior draws. Put differently, model uncertainty is pervasive and the best prediction is obtained as a weighted average of several models, each including between 20 and 50 regressors, but not necessarily the same ones across models. These results explain the empirical success of Bayesian model averaging techniques, such as those of Wright (2009), Faust et al. (2013), Fernandez et al. (2001), Sala-I-Martin et al. (2004), Cremers (2002) and Avramov (2002).

4.4. **Patterns of sparsity with an informative prior on model size.** The previous subsections suggest that the data are quite informative about the choice of shrinkage and model size, but the specific set of predictors to include in the model is generally not well identified. As a consequence, our posterior results are typically not compatible with the existence of a clear sparse representation of the predictive model, if the analysis is based on an agnostic prior on model size. This section shows that clearer patterns of sparsity can emerge when the researcher has a strong a-priori bias in favor of predictive models with a small number of regressors.

To demonstrate this point, we replace our uniform hyperprior on q with a $Beta\,(1,k)$. This particular formulation is appealing because it implies an approximately exponential density for the number of included predictors. Specifically, the probability that a model of size qk is smaller than a given number s converges to $1-e^{-s}$ when the number of predictors k is large, as in our applications. The resulting hyperprior is heavily skewed to the right,

attributing a probability of five percent to models with more than three predictors, and less than one percent to models with five or more predictors.[8]

The posterior probabilities of inclusion obtained with this tight hyperprior are reported in figure 4.6. Relative to our baseline, these heat maps exhibits more white areas, revealing clearer patterns of sparsity in all six applications. For example, in the case of macro 1, there are now three distinct groups of predictors, those that always belong to the model, those included with positive probability and those never selected. A similar interpretation applies to the other applications, with the exception of finance 2, in which systematically excluded regressors are still rare.

We inspect the mechanism underlying these results by analyzing in greater detail the estimates obtained with the flat and tight hyperprior in the macro-1 case. The intuition is similar in the other applications. The heat map in the top-left panel of figure 4.7 visualizes the probability of inclusion of each predictor, conditional on a specific value of the overall probability of inclusion q. The top-right panel plots instead the posterior of q, along with our baseline uniform prior. Observe that high-density values of q are mostly associated with non-white regions in the heat map, confirming the main takeaway of section 4.3 about the absence of clear sparsity patterns. However, a few stripes are very dark from top to bottom, indicating that the corresponding regressors are always included in the model, regardless of model size. When q is very low, these predictors are the only ones to be selected. Therefore, if the posterior of q were concentrated on these very low values, contrary to our findings, the results would be more consistent with the presence of sparsity.

The bottom-right panel of figure 4.7 presents the posterior density of q when the model is estimated with the tight $Beta\,(1, k)$ hyperprior depicted by the solid line. Not surprisingly, such a dogmatic prior generates a posterior concentrated on much smaller values of q, with essentially zero mass on values larger than 0.15. As a consequence, the corresponding heat map is mostly empty, except in the region of very low q's, where the heat map is nearly identical to that obtained with the flat hyperprior. Hence, in practice, the tight hyperprior effectively implements a censoring of the heat map, and simply restricts the inference to a region characterized by a higher degree of sparsity.

[8]This type of dogmatic priors is used by Castillo et al. (2015) to establish sparsistency, i.e. the ability to select the correct sparse model if it exists.

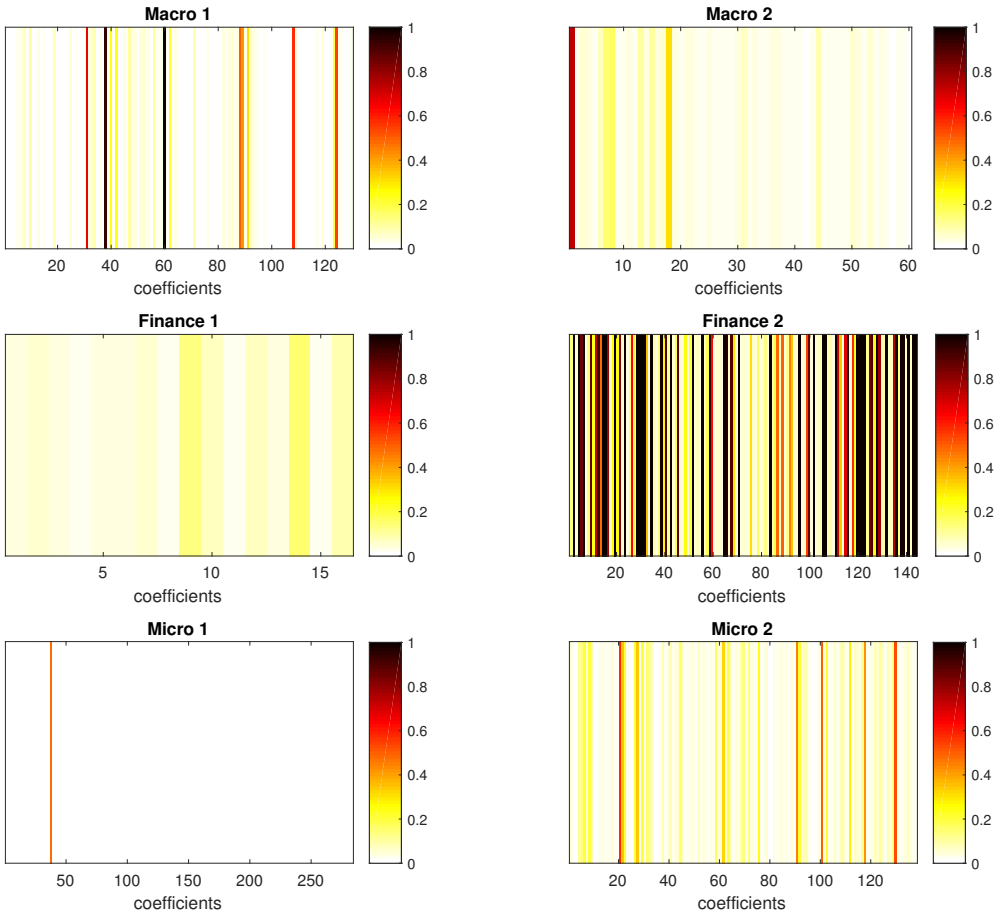

FIGURE 4.6. Heat map of the probabilities of inclusion of each predictor with a tight prior on q (best viewed in color).

Summing up, strong prior beliefs favoring low dimensional models appear to be necessary to compensate for the lower fit of these models documented in section 4.2, and thus to support sparse representations. This is why we conclude that the idea that the data are informative enough to identify sparse predictive models might just be an illusion in most cases.

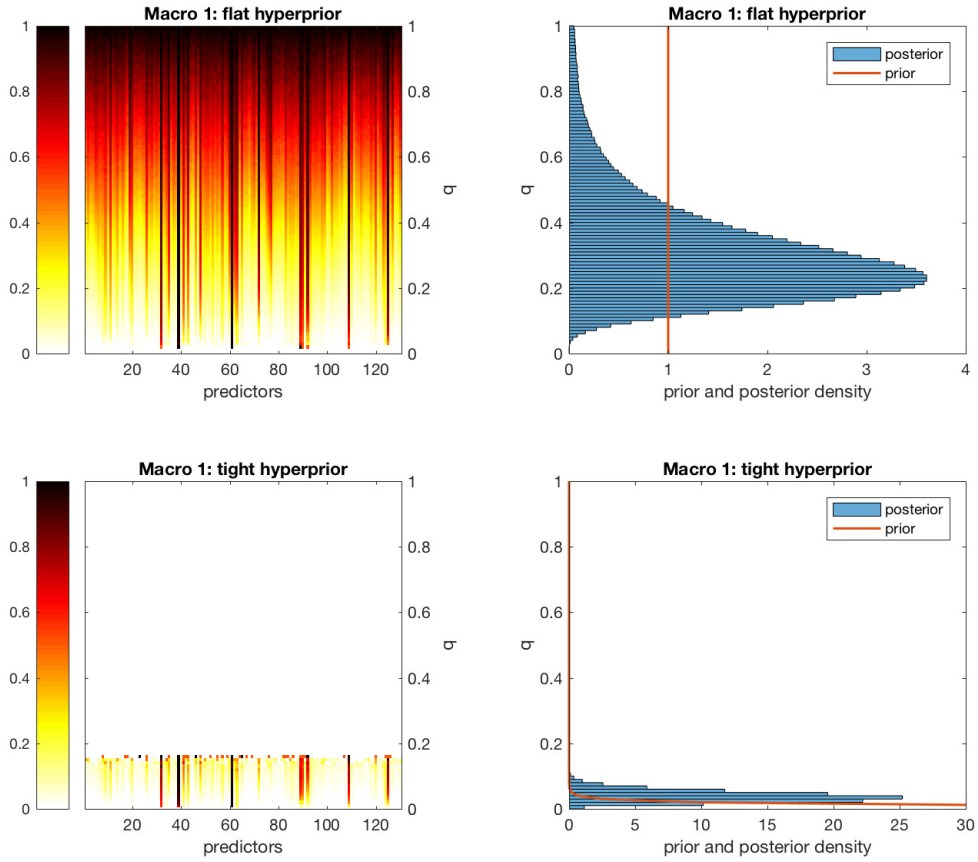

FIGURE 4.7. Heat map of the probabilities of inclusion of each predictor, given q (left panels), and prior and posterior densities of q (right panels), with a flat and tight prior on q in the macro-1 application (best viewed in color).

APPENDIX A. ALGORITHM FOR POSTERIOR INFERENCE

To estimate the model, it is useful to rewrite it using a set of latent variables $z = [z_1, ..., z_k]'$ that are equal to 1 when the corresponding regressor is included in the model and its coefficient is non-zero. Let us denote by $Y = [y_1, ..., y_T]'$, $U = [u_1, ..., u_T]'$ and $X = [x_1, ..., x_T]'$, where T is the number of observations. The posterior of the unknown objects of the model is given by

$$p\left(\phi, \beta, \sigma^2, R^2, z, q|Y, X\right) \propto p\left(Y|X, \phi, \beta, \sigma^2, R^2, z, q\right) \cdot p\left(\phi, \beta, \sigma^2, R^2, z, q\right)$$

$$\propto p\left(Y|X, \phi, \beta, \sigma^2\right) \cdot p\left(\beta|\sigma^2, R^2, z, q\right) \cdot p\left(z|q, \sigma^2, R^2\right) \cdot p\left(q\right) \cdot p\left(\sigma^2\right) \cdot p\left(R^2\right)$$

$$\propto \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{T}{2}} e^{-\frac{1}{2\sigma^2}(Y-U\phi-X\beta)'(Y-U\phi-X\beta)}$$
$$\cdot \prod_{i=1}^{k}\left[\left(\frac{1}{2\pi\sigma^2\gamma^2}\right)^{\frac{1}{2}} e^{-\frac{\beta_i^2}{2\sigma^2\gamma^2}}\right]^{1-z_i} \left[\delta\left(\beta_i\right)\right]^{z_i}$$
$$\cdot \prod_{i=1}^{k} q^{1-z_i}\left(1-q\right)^{z_i}$$
$$\cdot q^{a-1}\left(1-q\right)^{b-1}$$
$$\cdot \left(\frac{1}{\sigma^2}\right)$$
$$\cdot \left(R^2\right)^{A-1}\left(1-R^2\right)^{B-1},$$

where $\delta(\cdot)$ is the Dirac-delta function, which is the limit of $\left(\frac{1}{2\pi\alpha^2}\right)^{\frac{1}{2}} e^{-\frac{\beta_i^2}{2\alpha^2}}$ for $\alpha \to 0$.

We can sample from the posterior of $\left(\phi, \beta, \sigma^2, R^2, z, q\right)$ using a Gibbs sampling algorithm with blocks (i) R^2 and q, (ii) ϕ, and (iii) $\left(z, \beta, \sigma^2\right)$.

- The conditional posterior of R^2 and q is given by

$$p\left(R^2, q|Y, \phi, \beta, \sigma^2, z\right) \propto \left[e^{-\frac{1}{2\sigma^2}\frac{k\ varx\ q\left(1-R^2\right)}{R^2}\beta' diag(z)\beta}\right] \cdot$$
$$\cdot q^{\tau(z)+\frac{\tau(z)}{2}+a-1}\left(1-q\right)^{k-\tau(z)+b-1} \cdot \left(R^2\right)^{A-1-\frac{\tau(z)}{2}}\left(1-R^2\right)^{\frac{\tau(z)}{2}+B-1},$$

  where $\tau(z) \equiv \sum_{i=1}^{k} z_i$. We can sample from this distribution by discretizing the $[0,1]$ support of R^2 and q. More specifically, for both R^2 and q we define a grid with increments of 0.01, and finer increments of 0.001 near the boundaries of the support.

- The conditional posterior of ϕ is given by

$$p\left(\phi|Y, z, \beta, R^2, q, \sigma\right) \propto e^{-\frac{1}{2\sigma^2}(Y-U\phi-X\beta)'(Y-U\phi-X\beta)},$$

  which implies

$$\phi|Y, z, \beta, \gamma, q, \sigma \sim N\left(\left(U'U\right)^{-1}U'\left(Y-X\beta\right), \sigma^2\left(U'U\right)^{-1}\right).$$

- To draw from the posterior of $z, \beta, \sigma^2|Y, \phi, R^2, q$, we will first draw from $p\left(z|Y, \phi, R^2, q\right)$, and then from $p\left(\beta, \sigma^2|Y, \phi, R^2, q, z\right)$. To draw from the posterior $z|Y, \phi, R^2, q$, observe that‘

$$p\left(z|Y, \phi, R^2, q\right) = \int p\left(z, \beta, \sigma^2|Y, \phi, R^2, q\right) d\left(\beta, \sigma^2\right)$$

$$\propto q^{\tau(z)} (1-q)^{k-\tau(z)} \left(\frac{1}{2\pi\gamma^2}\right)^{\frac{\tau(z)}{2}} \int \left(\frac{1}{\sigma^2}\right)^{\frac{T+\tau(z)}{2}+1} e^{-\frac{1}{2\sigma^2}\left[\left(Y-U\phi-\tilde{X}\tilde{\beta}\right)'\left(Y-U\phi-\tilde{X}\tilde{\beta}\right)+\tilde{\beta}'\tilde{\beta}/\gamma^2\right]} d\left(\tilde{\beta}, \sigma^2\right)$$

$$\propto q^{\tau(z)} (1-q)^{k-\tau(z)} \left(\frac{1}{2\pi\gamma^2}\right)^{\frac{\tau(z)}{2}} (2\pi)^{\frac{\tau(z)}{2}} \left|\tilde{W}\right|^{-\frac{1}{2}} \int \left(\frac{1}{\sigma^2}\right)^{\frac{T}{2}+1} e^{-\frac{1}{2\sigma^2}\left[\tilde{Y}'\tilde{Y}-\hat{\tilde{\beta}}'\tilde{W}\hat{\tilde{\beta}}\right]} d\sigma^2$$

$$\propto q^{\tau(z)} (1-q)^{k-\tau(z)} \left(\frac{1}{\gamma^2}\right)^{\frac{\tau(z)}{2}} \left|\tilde{W}\right|^{-\frac{1}{2}} \left[\frac{\tilde{Y}'\tilde{Y}-\hat{\tilde{\beta}}'\tilde{W}\hat{\tilde{\beta}}}{2}\right]^{-\frac{T}{2}} \Gamma\left(\frac{T}{2}\right),$$

where $\gamma^2 = \frac{1}{k\ varx\ q} \cdot \frac{R^2}{1-R^2}$, $\tilde{\beta}$ is the vector of the non-zero coefficients (i.e. those corresponding to $z_i = 1$), $\tilde{X}$ are the corresponding regressors, $\hat{\tilde{\beta}} = \tilde{W}^{-1}\tilde{X}'\tilde{Y}$, $\tilde{W} = \left(\tilde{X}'\tilde{X} + I_{\tau(z)}/\gamma^2\right)$, and $\tilde{Y} = Y - U\phi$. Therefore, to draw from the posterior of $z|Y, \phi, R^2, q$, we can use a Gibbs sampler that allows to draw from the distribution of $z_i|Y, \phi, R^2, q, z_{-i}$. Finally, to draw from the posterior of $\beta, \sigma^2|Y, \phi, R^2, q, z$, observe that

$$\sigma^2|Y, \phi, R^2, q, z \sim IG\left(\frac{T}{2}, \frac{\tilde{Y}'\tilde{Y} - \hat{\tilde{\beta}}'\left(\tilde{X}'\tilde{X} + I_{\tau(z)}/\gamma^2\right)\hat{\tilde{\beta}}}{2}\right)$$

and

$$\tilde{\beta}|Y, \phi, \sigma^2, R^2, q, z \sim N\left(\hat{\tilde{\beta}}, \sigma^2\left(\tilde{X}'\tilde{X} + I_{\tau(z)}/\gamma^2\right)^{-1}\right)$$

and the other β_i's are equal to 0.

References

Avramov, D. (2002): "Stock return predictability and model uncertainty," *Journal of Financial Economics*, 64, 423 – 458.

Barro, R. J. (1991): "Economic Growth in a Cross Section of Countries," *The Quarterly Journal of Economics*, 106, 407–443.

Barro, R. J. and J.-W. Lee (1994): "Sources of economic growth," *Carnegie-Rochester Conference Series on Public Policy*, 40, 1 – 46.

Belloni, A., D. L. Chen, V. Chernozhukov, and C. Hansen (2012): "Sparse models and methods for optimal instruments with an application to eminent domain," *Econometrica*, 80, 2369–2429.

Belloni, A., V. Chernozhukov, and C. Hansen (2011): "Inference for high-dimensional sparse econometric models," in *Advances in Economics and Econometrics – World Congress of Econometric Society 2010*.

——— (2014): "Inference on Treatment Effects after Selection among High-Dimensional Controls," *The Review of Economic Studies*, 81, 608.

Carvalho, C. M., N. G. Polson, and J. G. Scott (2010): "The horseshoe estimator for sparse signals," *Biometrika*, 97, 465–480.

Castillo, I., J. Schmidt-Hieber, and A. van der Vaart (2015): "Bayesian linear regression with sparse priors," *Annals of Statistics*, 43, 1986–2018.

Chen, D. L. and S. Yeh (2012): "Growth under the shadow of expropriation? The economic impacts of eminent domain," Mimeo, Toulouse School of Economics.

Cremers, K. M. (2002): "Stock return predictability: A Bayesian model selection perspective," *Review of Financial Studies*, 15, 1223–1249.

De Mol, C., D. Giannone, and L. Reichlin (2008): "Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components?" *Journal of Econometrics*, 146, 318–328.

Donohue, J. J. and S. D. Levitt (2001): "The impact of legalized abortion on crime," *The Quarterly Journal of Economics*, 116, 379–420.

Faust, J., S. Gilchrist, J. H. Wright, and E. Zakrajsek (2013): "Credit Spreads as Predictors of Real-Time Economic Activity: A Bayesian Model-Averaging Approach," *The Review of Economics and Statistics*, 95, 1501–1519.

Fernandez, C., E. Ley, and M. F. J. Steel (2001): "Model uncertainty in cross-country growth regressions," *Journal of Applied Econometrics*, 16, 563–576.

Freyberger, J., A. Neuhierl, and M. Weber (2017): "Dissecting Characteristics Nonparametrically," Working Paper 23227, National Bureau of Economic Research.

George, E. I. and D. P. Foster (2000): "Calibration and empirical Bayes variable selection," *Biometrika*, 87, 731–747.

George, E. I. and R. E. McCulloch (1993): "Variable Selection via Gibbs Sampling," *Journal of the American Statistical Association*, 88, 881–889.

——— (1997): "Approaches for Bayesian Variable Selection," *Statistica Sinica*, 7, 339–373.

Harvey, C. R., Y. Liu, and H. Zhu (2016): "...and the Cross-Section of Expected Returns," *The Review of Financial Studies*, 29, 5.

Hastie, T., R. Tibshirani, and M. Wainwright (2015): *Statistical learning with sparsity*, CRC press.

Hoerl, A. E. and R. W. Kennard (1970): "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, 12, 55–67.

Ishwaran, H. and J. S. Rao (2005): "Spike and slab variable selection: Frequentist and Bayesian strategies," *Ann. Statist.*, 33, 730–773.

Lawley, D. N. and A. E. Maxwell (1963): *Factor analysis as a statistical method [by] D.N. Lawley and A.E. Maxwell*, Butterworths London.

Leamer, E. E. (1973): "Multicollinearity: A Bayesian Interpretation," *The Review of Economics and Statistics*, 55, 371–380.

Liang, F., R. Paulo, G. Molina, M. A. Clyde, and J. O. Berger (2008): "Mixtures of g priors for Bayesian variable selection," *Journal of the American Statistical Association*, 103, 410–423.

McCracken, M. W. and S. Ng (2016): "FRED-MD: A Monthly Database for Macroeconomic Research," *Journal of Business & Economic Statistics*, 34, 574–589.

Mitchell, T. J. and J. J. Beauchamp (1988): "Bayesian Variable Selection in Linear Regression," *Journal of the American Statistical Association*, 83, 1023–1032.

Park, T. and G. Casella (2008): "The Bayesian Lasso," *Journal of the American Statistical Association*, 103, 681–686.

Pearson, K. (1901): "On lines and planes of closest fit to systems of points in space," *Philosophical Magazine Series 6*, 2, 559–572.

Sala-I-Martin, X., G. Doppelhofer, and R. I. Miller (2004): "Determinants of Long-Term Growth: A Bayesian Averaging of Classical Estimates (BACE) Approach," *American Economic Review*, 94, 813–835.

Spearman, C. (1904): ""General Intelligence," Objectively Determined and Measured," *The American Journal of Psychology*, 15, 201–292.

Stock, J. H. and M. W. Watson (2002a): "Forecasting Using Principal Components from a Large Number of Predictors," *Journal of the American Statistical Association*, 97, 147–162.

——— (2002b): "Macroeconomic Forecasting Using Diffusion Indexes," *Journal of Business and Economics Statistics*, 20, 147–162.

Tibshirani, R. (1996): "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Tikhonov, A. N. (1963): "Solution of Incorrectly Formulated Problems and the Regularization Method," *Soviet Math. Dokl.*, 5, 1035/1038.

Welch, I. and A. Goyal (2008): "A Comprehensive Look at The Empirical Performance of Equity Premium Prediction," *Review of Financial Studies*, 21, 1455–1508.

Wright, J. H. (2009): "Forecasting US inflation by Bayesian model averaging," *Journal of Forecasting*, 28, 131–144.

Federal Reserve Bank of New York and CEPR

E-mail address: dgiannon2@gmail.com

European Central Bank and ECARES

E-mail address: michele.lenza@ecb.int

Northwestern University, CEPR and NBER

E-mail address: g-primiceri@northwestern.edu